Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.01

Lecture 24: Continued Introduction to Bayesian Inference and Use of the MCMC Algorithm for Inference

Jason Mezey
jgm45@cornell.edu
May 1, 2018 (T) 8:40-9:55
Announcements

• Note two room changes (!!): May 8 changed to BB 302-C

• Reminder that your project is due (11:59PM) Tues., next week = May 8 (last day of class!)

• Final Exam:
  • Available Mon., May 14, Due Fri. (11:59PM) May 18
  • Open book / take home, same format / rules as midterm
  • Cumulative
Review: Intro to Bayesian analysis I

- Remember that in a Bayesian (not frequentist!) framework, our parameter(s) have a probability distribution associated with them that reflects our belief in the values that might be the true value of the parameter.

- Since we are treating the parameter as a random variable, we can consider the joint distribution of the parameter AND a sample $Y$ produced under a probability model:

$$Pr(\theta \cap Y)$$

- For inference, we are interested in the probability the parameter takes a certain value given a sample:

$$Pr(\theta | y)$$

- Using Bayes theorem, we can write:

$$Pr(\theta | y) = \frac{Pr(y | \theta)Pr(\theta)}{Pr(y)}$$

- Also note that since the sample is fixed (i.e. we are considering a single sample) $Pr(y) = c,$ we can rewrite this as follows:

$$Pr(\theta | y) \propto Pr(y | \theta)Pr(\theta)$$
Review: Intro to Bayesian analysis II

• Let’s consider the structure of our main equation in Bayesian statistics:

\[ Pr(\theta|y) \propto Pr(y|\theta)Pr(\theta) \]

• Note that the left hand side is called the posterior probability:

\[ Pr(\theta|y) \]

• The first term of the right hand side is something we have seen before, i.e. the likelihood (!!):

\[ Pr(y|\theta) = L(\theta|y) \]

• The second term of the right hand side is new and is called the prior:

\[ Pr(\theta) \]

• Note that the prior is how we incorporate our assumptions concerning the values the true parameter value may take.

• In a Bayesian framework, we are making two assumptions (unlike a frequentist where we make one assumption: 1. the probability distribution that generated the sample, 2. the probability distribution of the parameter.
• Let’s put this all together for our “heights in the US” example

• First recall that our assumption is the probability model is normal (so what is the form of the likelihood?):

\[ Y \sim N(\mu, \sigma^2) \]

• Second, assume a normal prior for the parameter we are interested in:

\[ Pr(\mu) \sim N(\kappa, \phi^2) \]

• From the Bayesian equation, we can now put this together as follows:

\[ Pr(\theta|y) \propto Pr(y|\theta)Pr(\theta) \]

\[ Pr(\mu|y) \propto \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \right) \frac{1}{\sqrt{2\pi\phi^2}} e^{-\frac{(\mu-\kappa)^2}{2\phi^2}} \]

• Note that with a little rearrangement, this can be written in the following form:

\[ Pr(\mu|y) \sim N\left( \left( \frac{\kappa}{\sigma^2} + \frac{\sum_{i}^{n} y_i}{\sigma^2} \right), \left( \frac{1}{\phi^2} + \frac{n}{\sigma^2} \right)^{-1} \right) \]
Review: Bayesian inference

• Inference in a Bayesian framework differs from a frequentist framework in both estimation and hypothesis testing.

• For example, for estimation in a Bayesian framework, we always construct estimators using the posterior probability distribution, for example:

\[
\hat{\theta} = mean(\theta | y) = \int \theta Pr(\theta | y) d\theta \quad \text{or} \quad \hat{\theta} = median(\theta | y)
\]

• For example, in our “heights in the US” example our estimator is:

\[
\hat{\mu} = median(\mu | y) = mean(\mu | y) = \frac{\left(\frac{\kappa}{\sigma^2} + \frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\phi^2} + \frac{n}{\sigma^2}\right)}
\]

• For hypothesis testing we could (and most appropriately) use Bayes factor, although in this class and in many cases in practice we will use a “psuedo-Bayesian” approach were we assess if the credible interval (e.g. the 0.95 c.i.) of the posterior distribution overlaps the value of the parameter under the null hypothesis.

• Estimates in a Bayesian framework can be different than in a likelihood (Frequentist) framework since estimator construction is fundamentally different (!!!)
We are now ready to tackle Bayesian inference for our genetic model (note that we will focus on the linear regression model but we can perform Bayesian inference for any GLM!): 

\[ Y = \beta_{\mu} + X_a \beta_a + X_d \beta_d + \epsilon \]

\[ \epsilon \sim N(0, \sigma_\epsilon^2) \]

Recall for a sample generated under this model, we can write:

\[ y = x \beta + \epsilon \]

\[ \epsilon \sim multiN(0, I\sigma_\epsilon^2) \]

In this case, we are interested in the following hypotheses:

\[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]

\[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]

We are therefore interested in the marginal posterior probability of these two parameters.
Review: a genetic model II

- To calculate these probabilities, we need to assign a joint probability distribution for the prior

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) \]

- One possible choice is as follows (are these proper or improper!?):

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) = Pr(\beta_\mu) Pr(\beta_a) Pr(\beta_d) Pr(\sigma^2_\epsilon) \]

\[ Pr(\beta_\mu) = Pr(\beta_a) = Pr(\beta_d) = c \]

\[ Pr(\sigma^2_\epsilon) = c \]

- Under this prior the complete posterior distribution is multivariate normal (!!):

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon | y) \propto Pr(y | \beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) \]

\[ Pr(\theta | y) \propto (\sigma^2_\epsilon)^{-\frac{n}{2}} e^{-\frac{(y-\mathbf{x}\beta)^T(y-\mathbf{x}\beta)}{2\sigma^2_\epsilon}} \]
Review: a genetic model III

- For the linear model with sample:

\[ y = x\beta + \epsilon \]

\[ \epsilon \sim \text{multi}N(0, I\sigma_\epsilon^2) \]

- The complete posterior probability for the genetic model is:

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | y) \propto Pr(y | \beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2) Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2) \]

- With a uniform prior is:

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | y) \propto Pr(y | \beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2) \]

- The marginal posterior probability of the parameters we are interested in is:

\[ Pr(\beta_a, \beta_d | y) = \int_0^\infty \int_{-\infty}^\infty Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | y) d\beta_\mu d\sigma_\epsilon^2 \]
Review: a genetic model IV

- Assuming uniform (improper!) priors, the marginal distribution is:

\[
Pr(\beta_a, \beta_d | y) = \int_{-\infty}^{\infty} \int_{0}^{\infty} Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | y) d\beta_\mu d\sigma_\epsilon^2 \sim \text{multi-t-distribution}
\]

- With the following parameter values:

\[
\text{mean}(Pr(\beta_a, \beta_d | y)) = \begin{bmatrix} \hat{\beta}_a, \hat{\beta}_d \end{bmatrix}^T = C^{-1} [X_a, X_d]^T y \quad \quad C = \begin{bmatrix} X_a^T X_a & X_a^T X_d \\ X_d^T X_a & X_d^T X_d \end{bmatrix}
\]

\[
cov = \frac{(y - [X_a, X_d] \begin{bmatrix} \hat{\beta}_a, \hat{\beta}_d \end{bmatrix}^T)^T (y - [X_a, X_d] \begin{bmatrix} \hat{\beta}_a, \hat{\beta}_d \end{bmatrix}^T)}{n - 6} C^{-1}
\]

\[
df (\text{multi-t}) = n - 4
\]

- With these estimates (equations) we can now construct a credible interval for our genetic null hypothesis and test a marker for a phenotype association and we can perform a GWAS by doing this for each marker (!!)

November 20, 2012
Review: Bayesian “hypothesis testing”

$\Pr(\beta_a, \beta_d | y)$

Cannot reject H0!

$\Pr(\beta_a, \beta_d | y)$

Reject H0!
Bayesian inference for more “complex” posterior distributions

- For a linear regression, with a simple (uniform) prior, we have a simple closed form of the overall posterior.

- This is not always (=often not the case), since we may often choose to put together more complex priors with our likelihood or consider a more complicated likelihood equation (e.g. for a logistic regression!)

- To perform hypothesis testing with these more complex cases, we still need to determine the credible interval from the posterior (or marginal) probability distribution so we need to determine the form of this distribution.

- To do this we will need an algorithm and we will introduce the Markov chain Monte Carlo (MCMC) algorithm for this purpose.
Review: Stochastic processes

- To introduce the MCMC algorithm for our purpose, we need to consider models from another branch of probability (remember, probability is a field much larger than the components that we use for statistics / inference!): Stochastic processes

- **Stochastic process** (intuitive def) - a collection of random vectors (variables) with defined conditional relationships, often indexed by a ordered set $t$

- We will be interested in one particular class of models within this probability sub-field: *Markov processes* (or more specifically *Markov chains*)

- Our MCMC will be a Markov chain (probability model)
Review: Markov processes

- A Markov chain can be thought of as a random vector (or more accurately, a set of random vectors), which we will index with $t$:

  \[ X_t, X_{t+1}, X_{t+2}, \ldots, X_{t+k} \]

  \[ X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-k} \]

- Markov chain - a stochastic process that satisfies the Markov property:

  \[ Pr(X_t | X_{t-1}, X_{t-2}, \ldots, X_{t-k}) = Pr(X_t | X_{t-1}) \]

- While we often assume each of the random variables in a Markov chain are in the same class of random variables (e.g. Bernoulli, normal, etc.) we allow the parameters of these random variables to be different, e.g. at time $t$ and $t+1$

- How does this differ from a random vector of an iid sample!?
Review: example of a Markov chain

- As an example, let’s consider a Markov chain where each random variable in the chain has a Bernoulli distribution:

\[ X_1, X_2, \ldots, X_{1001}, X_{1002} \]

\[ X_1 \sim Bern(0.2), X_2 \sim Bern(0.45), \ldots, X_{1001} \sim Bern(0.4), X_{1002} \sim Bern(0.4) \]

- Note that we could draw observations from this Markov chain (since it is just a random vector with a probability distribution!):

\[ 1,0,\ldots,1,1 \quad 0,0,\ldots,0,0 \]

\[ 0,1,\ldots,1,1 \quad 0,1,\ldots,0,0 \]

- How does this differ from an iid random vector?

- Note that for \( t \) late in this process, the parameters of the Bernoulli distributions are the same (=they do not change over time)

- In our case, we will be interested in Markov chains that “evolve” to such stationary distributions
Stationary distributions and MCMC

- If a Markov chain has certain properties (irreducible and ergodic), we can prove that the chain will evolve (more accurately converge!) to a unique (!!) stationary distribution and will not leave this stationary distribution (where is it often possible to determine the parameters for the stationary distribution!)

- For such Markov chains, if we consider enough iterations $t+k$ (where $k$ may be very large, e.g. infinite), we will reach a point where each following random variable is in the unique stationary distribution:

  $$Pr(X_{t+k}) = Pr(X_{t+k+1}) = \ldots$$

- For the purposes of Bayesian inference, we are going to set up a Markov chain that evolves to a unique stationary distribution that is exact**ly** the posterior probability distribution that we are interested in (!!!)

- To use this chain, we will run the Markov chain for enough iterations to reach this stationary distribution and then we will take a sample from this chain to determine (or more accurately approximate) our posterior

- This is Bayesian Markov chain Monte Carlo (MCMC)!
An example of Bayesian MCMC

\[ MCMC = X_{t+k}, X_{t+k+1}, X_{t+k+2}, \ldots, X_{t+k+m} \]

Sample = 0.1, −0.08, −1.4, ..., 0.5

\[ \hat{\theta} = \text{median}(Pr(\theta|y)) \approx \text{median}(\theta^{[tab]}, \ldots, \theta^{[tab+k]}) \]
Constructing an MCMC

- Instructions for constructing an MCMC using Metropolis-Hastings approach:

  1. Choose $\theta^0$, where $Pr(\theta^0|y) > 0$.
  2. Sample a proposal parameter value $\theta^*$ from a jumping distribution $J(\theta^*|\theta^t)$, where $t = 0$ or any subsequent iteration.
  3. Calculate $r = \frac{Pr(\theta^*|y)J(\theta^t|\theta^*)}{Pr(\theta^t|y)J(\theta^*|\theta^t)}$.
  4. Set $\theta^{t+1} = \theta^*$ with $Pr(\theta^{t+1} = \theta^*) = \min(r, 1)$ and $\theta^{t+1} = \theta^t$ with $Pr(\theta^{t+1} = \theta^t) = 1 - \min(r, 1)$.

- Running the MCMC algorithm:

  1. Set up the Metropolis-Hastings algorithm.
  2. Initialize the values for $\theta^0$.
  3. Iterate the algorithm for $t >> 0$, such that we are past $t_{ab}$, which is the iteration after the ‘burn-in’ phase, where the realizations of $\theta^t$ start to behave as though they are sampled from the stationary distribution of the Metropolis-Hastings Markov chain (we will discuss how many iterations are necessary for a burn-in below).
  4. Sample the chain for a set of iterations after the burn-in and use these to approximate the posterior distribution and perform Bayesian inference.
That’s it for today

• We will introduce pedigree and inbred line analysis AND the basics of classic quantitative genetics (additive genetic variance and heritability) in the last two lectures!